PDQ Wizard: automated prioritization and characterization of gene and protein lists using biomedical literature

BIOINFORMATICS APPILICATIONS NOTE Vol. 22 no. 16 2006, pages 2055–2057
doi:10.1093/bioinformatics/btl342

Data and text mining

PDQ Wizard: automated prioritization and characterization of gene and protein lists using biomedical literature

G. R. Grimes[1,*], T. Q. Wen[2], M. Mewissen[1], R. M. Baxter[3], S. Moodie[1], J. S. Beattie[1] and P. Ghazal[1]

[1]The Scottish Centre for Genomic Technology and Informatics, University of Edinburgh, 49 Little France Crescent, Edinburgh EH16 4SB, UK, [2]eDIKT Programme, National E-Science Centre, 15 South College Street, Edinburgh EH8 9AA, UK and [3]Edinburgh Parallel Computing Centre, The University of Edinburgh, James Clerk Maxwell Building, King's Buildings, Edinburgh EH9 3JZ, UK

Received on March 16, 2006; revised on May 10, 2006; accepted on June 20, 2006
Advance Access publication June 29, 2006
Associate Editor: Thomas Lengauer

ABSTRACT

Summary: PDQ Wizard automates the process of interrogating biomedical references using large lists of genes, proteins or free text. Using the principle of linkage through co-citation biologists can mine PubMed with these proteins or genes to identify relationships within a biological field of interest. In addition, PDQ Wizard provides novel features to define more specific relationships, highlight key publications describing those activities and relationships, and enhance protein queries. PDQ Wizard also outputs a metric that can be used for prioritization of genes and proteins for further research.

Availability: PDQ Wizard is freely available from http://www.gti.ed.ac.uk/pdqwizard/

Contact: Graeme.Grimes@ed.ac.uk

Supplementary Information: Supplementary Data are available http://www.gti.ed.ac.uk/pdqwizard/

INTRODUCTION

High-throughput technologies are now widely used for the global and parallel measurement of gene and protein activity within biological systems. A primary output from these analyses is often a collection of tens or hundreds of genes or proteins of interest. A major challenge for biologists, therefore, is to rapidly derive comprehensive information about the biological processes for each of the specific genes or proteins in the list and to identify where domain-specific relationships exist. Several databases, such as Entrez Gene (Maglott *et al*., 2005) and UniProt (Bairoch *et al*., 2005) enable biologists to access information on individual genes and proteins. Biologists, however, frequently require more in-depth, specific information than is included in these databases and need to be able to explore gene and protein lists rather than individual identifiers.

The detailed information biologists require is primarily stored as free text within large biomedical literature databases such as PubMed (Wheeler *et al*., 2005) which contains over 15 million references. Significantly, Entrez (Wheeler *et al*., 2005) which is the main interface for searching and retrieving information from PubMed, is not designed for searching with multiple gene or protein identifiers, such as Entrez Gene Ids. Consequently, it is inadequate for the rapid interrogation of literature relating to multiple genes and proteins. More generally, common descriptor terms such as gene symbols are insufficient for searching of the literature, owing to the fact that most genes are represented by multiple synonyms (Pearson, 2001). Therefore, there is a requirement for the inclusion of comprehensive annotations in order to retrieve all relevant information existing within literature resources.

Several tools, such as microGenie (Korotkiy *et al*., 2004) and MILANO (Rubinstein and Simon, 2005) have been developed to automate the annotation, batch query and data retrieval steps during PubMed searches. These gene-based search applications are limited to providing a single method to identify co-citation relationships, and they are restricted from further refinement of results or alternative querying strategies and do not permit the use of protein identifiers. For these reasons, we have sought to provide more flexible querying approaches and offer enhanced support for other types of high-throughput data.

PDQ Wizard provides a system that identifies relationships between lists of gene or protein identifiers and user defined terms based on their co-occurrence within PubMed literature references. The system outputs a table that includes the original gene or protein identifiers, with associated information such as the gene synonyms, gene description and the list of user defined terms. For each gene/protein Id and user defined term pair the number of PubMed records co-citing these terms are also displayed. Significantly, PDQ Wizard provides several novel features including the following:

- Interactive filtering of results, giving the ability to refine pairwise relationships and metrics for prioritization;
- Identification of top publications for a list of genes or proteins;
- Provides a view of publication information, including title and abstract, with syntax highlighting, similar to PubMed;
- Protein identifier input, providing support for Swiss-Prot identifiers.

*To whom correspondence should be addressed.

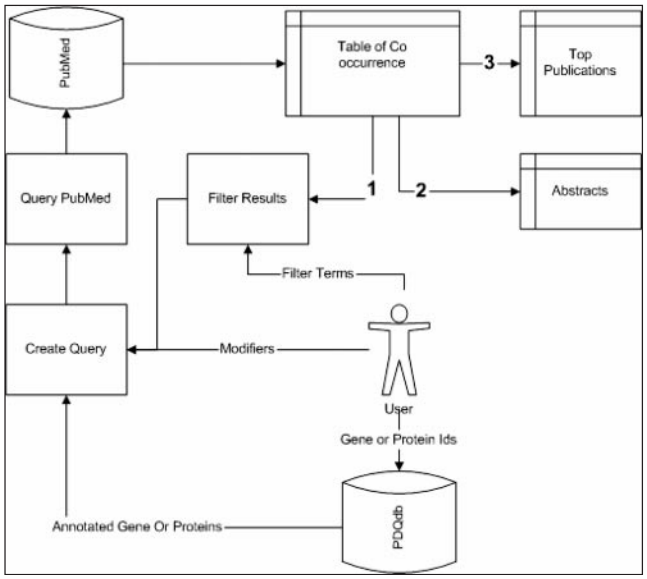

Fig. 1. PDQ Wizard work flow: The user enters a list of genes or proteins alongside a set of keyword terms. PDQ automatically annotates lists, generates PubMed queries and retrieves results. The results are presented as a table showing the number of co-citations for gene/protein identifier and user defined term pairs. The user has the choice of (1) Filtering results, (2) examining the references and (3) identifying publications that are present in multiple hits.

RESULTS AND DISCUSSION

PDQ Wizard was developed following a requirements capture process with biologists who regularly conduct manual literature searches involving large numbers of genes and proteins. Feedback from the users was used to enhance the usability and functionality of the system.

To cope with the multiplicity in biological naming, PDQ Wizard utilizes a gene and protein thesaurus derived from information stored within the UniProt and Entrez Gene databases. This is used to annotate identifiers with their corresponding official gene symbols, protein names, gene descriptions and synonyms. These annotations are automatically combined with user defined terms to construct enhanced PubMed queries. To limit the number of results retrieved due to synonymous terms within the literature, the thesaurus has been filtered to remove gene/protein synonyms that match words found within an English dictionary, biological acronyms and biological abbreviations. Gene names are not subject to filtering, however, they must match the exact phrase for a search to retrieve results. For example, for the *Drosophila* gene 'bag of marbles' the entire gene name must appear in the publication to classify as a hit.

In a typical example (Fig. 1), a biologist inputs a list of differentially regulated genes from a microarray experiment alongside a number of terms. These user defined terms are normally related to the biologist's field of scientific interest or the experimental system the lists are derived from. For example, for a list of differentially regulated genes derived from a microarray experiment where cells had been treated with interferon, a biologist may enter the term 'interferon'. Next PDQ Wizard queries PubMed and presents the results as a table of the pairwise co-occurrence of each gene or protein identifier and user defined term within PubMed. A 'hit' between an identifier and keyword indicates that both terms are co-cited within a PubMed record and may have an underlying relationship. Therefore, the user can use the finding of hits to categorize their list according to the relationship with keyword terms. The greater the number of hits, the more likely the inferred association (Marcotte and Date, 2001). As a result, biologists can use the number of hits to prioritize their future literature research based on the most likely gene/protein and user defined term relationships within their field of interest.

Biologists wishing to further categorize their lists can use the filter toolbar to input additional terms. The filter toolbar appends additional terms to the query table using the 'AND' operator. Users can also restrict these searches to specific fields within a PubMed record, e.g. title. For example, if an initial search has identified a subset of genes that have a relationship with 'interferon', a user may enter the term 'JAK' in the filter toolbar to identify which of those genes are related to the JAK pathway. The results now show the table of hits for the gene list, 'interferon' and 'JAK' (Supplementary Material), which can then be used to re-classify the gene list.

Another key task biologists perform is to identify publications that describe the relationship between multiple members of their gene or protein lists. PDQ Wizard provides the option to identify these key publications in the results using the 'top publication' feature. A top publication is defined as one that appears in multiple hits, so it should contain information that links multiple members of the gene or protein list with the user defined terms. This feature is especially useful for identifying those publications that describe biological pathways.

IMPLEMENTATION

PDQ Wizard is implemented as a Java Server Faces web application utilizing Apache Tomcat as the web server. The component that provides access to the PubMed server works through the Entrez utilities web service (Wheeler *et al*., 2005). The PubMed web service imposes limitations on its usage; this includes a maximum of one query every 3 seconds (Korotkiy *et al*., 2004). Therefore, to perform a search using 10 gene/protein identifiers and 10 user defined terms or 100 queries would take ~5 min. The gene/protein thesaurus is stored within a MySQL database that contains gene and protein annotations parsed from Entrez Gene and UniProt database files using custom Python scripts. PubMed abstracts downloaded for manual inspection are cached locally to increase response time and reduce the load on the PubMed server.

CONCLUSION

PDQ Wizard is a web-based tool that enables the rapid classification and prioritization of large lists of gene and protein identifiers using the biomedical literature. The classification is based on the presence of genes or proteins and user defined terms within the literature, and the prioritization is based on the number of literature references retrieved for each identifier and user defined term pair. The system also provides novel features to further classify results, highlight relevant publications and manually inspect literature references. Future versions will

include the ability to mine other literature resources such as OMIM, GeneRif and Google Scholar. Other areas of research will focus on using natural language processing to automatically extract the semantics of relationships within the results and provide a confidence score.

ACKNOWLEDGEMENTS

The authors thank their colleagues at the GTI and collaborators Alan Pemberton, Varrie Oglivie, Elaine Marshall, Mathieu Blanc and Mick Rae for contributing to this resource. This work was supported by the eDKIT project, the Edinburgh Parallel Computing Centre, SHEFC, EU funded Network of Excellence 'Infobiomed'—Contract no.: 507585, Scottish Enterprise and the European Regional Development Fund. Funding to pay the Open Access publication charges was provided by Wellcome Trust.

Conflict of Interest: none declared.

REFERENCES

Bairoch,A. *et al*. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res*., **33**, D154–159.

Korotkiy,M. *et al*. (2004) A tool for gene expression based PubMed search through combining data sources. *Bioinformatics*, **20**, 1980–1982.

Maglott,D. *et al*. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*., **33**, D54–58.

Marcotte,E. and Date,S. (2001) Exploiting big biology: integrating large-scale biological data for function inference. *Brief Bioinform*., **2**, 363–374.

Pearson,H. (2001) Biology's name game. *Nature*, **411**, 631–632.

Rubinstein,R. and Simon,I. (2005) MILANO—custom annotation of microarray results using automatic literature searches. *BMC Bioinformatics*, **6**, 12.

Wheeler,D.L. *et al*. (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **33**, D39–D45.